AA11

by Christian S.k Aditya

Submission date: 06-Jan-2020 01:55PM (UTC+0700)
Submission ID: 1239478157
File name: B2.pdf (344.7K)
Word count: 2285
Character count: 12308

Seminar Nasional Teknologi dan Rekayasa (SENTRA) 2018
ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

DOCUMENT PLAGIARISM DETECTION USING DAMERAU LEVENSHTEIN ALGORITHM AND QUERY EXPANSION

Christian Sri Kusuma Aditya*[1], Vinna Rahmayanti Setyaning Nastiti[2]
Teknik Informatika, Universitas Muhammadiyah Malang

Kontak Person:
Christian Sri Kusuma Aditya
Jl. Raya Tlogomas 246 Malang, Telp/Fax 0341-464318 ext 247
Email: christianskaditya@umm.ac.id

Abstract

Plagiarism is considered a criminal act provided in the Law of the Republic of Indonesia Number 19 Year 2002 about Copyright, therefore plagiarism activities should be avoided. Computerized plagiarism detection needs to be done to reduce plagiarism against another people's work rapidly. One of the plagiarist's efforts in avoiding the existing plagiarism detection application is manipulating documents by replacing many words with synonyms. This research proposes a design of plagiarism detection application by considering synonyms of words in documents, so that it can recognize documents even though the content is different, but the context is same. From the experiments that has been conducted, by comparing Damerau Levenshtein and Levenshtein Algorithm, the similarity values are relative similar for the handling of test case overall. But the handling of the case documents for typographical errors, Damerau Levenshtein Distance can recognize plagiarism documents better indicated by a higher similarity value 77,32%. And after integrated with query expansion, it can get significant result by detecting synonym words, so the application can detect document plagiarism more quieter well

Keywords: *Damerau Levenshtein Distance, plagiarism document, string matching*

1. Introduction

Today's technological era makes it easier for us to get information freely through the internet world. That ease can certainly have a very positive impact on science, but in practice there is no denying the negative impact. In line with the increase in published papers, there are also increasingly widespread ethical violations, one of which is plagiarism. Manually assess whether an indicated plagiarism document will require a large amount of time and money. So, it needs a computerized system to detect whether a document includes plagiarism or not. In a previous study [1] a plagiarism detection system study was done using the Levenshtein Distance algorithm for the development of android smartphone software. Levenshtein Distance use of string-matching algorithm by finding the distance value between the two strings of operations performed are: (1) insertion operation, (2) deletion operation, and (3) a letter subtitution operation.

In other studies, about string matching, the addition of an operation to find distance values is done by observing the exchange of word positions (transposition) by Christanti for the Indonesian word spelling correction system using the Damerau-Levenshtein Distance algorithm [2]. The use of Damerau-Levenshtein gets better results than using Levenshtein because it is able to achieve more variations in spelling errors, and the time difference in processing time can be justified for increased accuracy.

In contrast to detecting plagiarism in the program source code that has clear and regular rules and writing, detecting plagiarism between two text documents is more difficult because human language is more dynamic which is constantly changing and has many modifying factors that also change (flexible). Query expansion is a technique to connect vocabulary gaps between keywords and documents [3]. Although different in content, the appearance of words that have similar meanings of documents can be calculated through synonym-set keywords.

Detecting plagiarism in the program source code that has clear regular rules and writing, detecting plagiarism between two text documents is more difficult because human language is more dynamic which is constantly changing and has many modifying factors that also change (flexible). Query expansion is a technique to connect vocabulary gaps between keywords and documents [3]. Although different in content, the appearance of words that have similar meanings of documents can be calculated through synonym-set keywords.

This study aims to develop a document plagiarism detection system that is not only capable of detecting based on the examination of words per word or content, but also checking the equality of

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

meaning or context in order to produce a system that is more precise and optimal in detecting documents with plagiarism.

2. Research Method

Computerized plagiarism detection needs to be done to reduce plagiarism on the work of others, therefore the aim of this research is to design an application to detect plagiarism by using the Damerau Levenshtein Distance and query expansion and can recognize the equivalent of words (synonyms) in order to increase the percentage of success the system in recognizing the similarity of a text document even though the word structure is different but meaning is same. Often found a works in a different writing structure, but when viewed and observed the content or meaning has a fairly high resemblance. Based on similarity proportions, plagiarism is divided into three categories, mild plagiarism (<30 percent), moderate (30 percent - 70 percent), and weight (> 70 percent) [4]. In general, this research method can be seen in Figure 1.

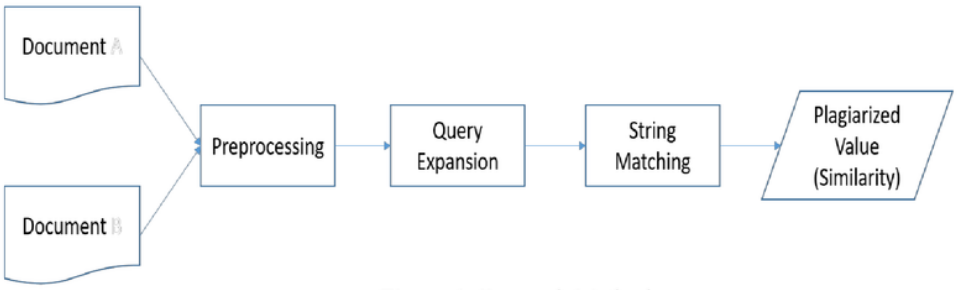


Figure 1. Research Methods

2.1. Data

For the test documents used in this study, there are several conditions used as follows:

1. doc-20%: is a test document which text content is deducted as much as 20% of the words randomly from the original document.
2. doc-40%: is a test document which text content is deducted as much as 40% of the words randomly from the original document.
3. doc-SE: is a test document which text content has exchanged the order of sentences randomly from the original document.
4. doc-AP: is a test document which verb structure is replaced to be active from passive or vice versa.
5. doc-syn: is a test document which words is turned into synonim that have same meaning from the original document.
6. doc-typo: is a test document with typographic errors from the original document.

2.2. System Analysis and Design

Documents that are compared in the system are documents abstraction section of a scientific work that will be compared with each other. Both documents will be carried out in each preprocessing stage. Preprocessing is done including case folding, tokenizing, stopword removal and stemming.

After going through preprocessing, the temporary output obtained is a collection of words that have undergone a change of form into a root word. This root word is then carried out a query expansion to get the synonym-set which will later expand the meaning of word that belongs to each string. This can overcome the possibility if the two strings are examined in a different structure or content but the context of the meaning of word is same, so the distance value when the string-matching process remains high. The following Figure 2 is modeling for the Query Expansion. Global model query expansion requires knowledge-based, in this study an Indonesian lexical database called Kateglo is used [5]. The total number of synonym pairs of words used 62912 records. The process of query expansion can be seen in Figure 2.

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

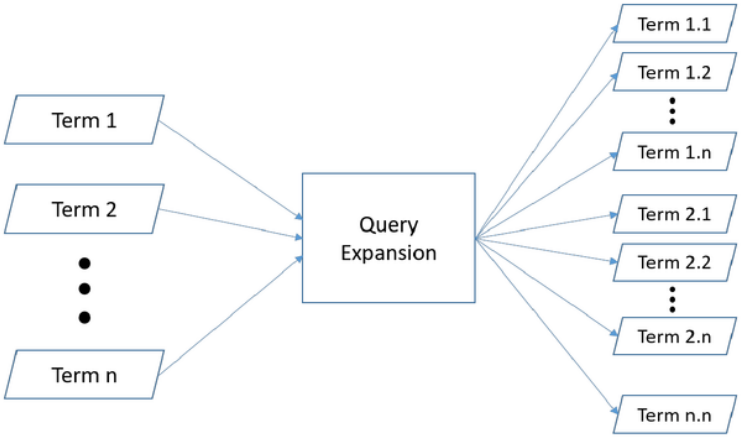


Figure 2. Query Expansion

Damerau Levenshtein Distance Algorithm is a development of Levenshtein Distance where there is the addition of a transposition operation of an adjacent letter. So that the obtained edit distance more optimal. Damerau Levenshtein Metric is a function in finite strings of an alphabet to integer. To express the Damerau–Levenshtein distance between two strings a and b a function $d_{a,b}(i,j)$ is defined, whose value is a distance between an i symbol prefix (initial substring) of a and j symbol prefix of b (Equation 1) [6].

$$d_{a,b}(i,j) = min\begin{cases} 0 & if\ i = j = 0 \\ d_{a,b}(i-1,j)+1 & if\ i > 0 \\ d_{a,b}(i,j-1)+1 & if\ j > 0 \\ d_{a,b}(i-1,j-1)+1_{(a_i \neq b_i)} & if\ i,j > 0 \\ d_{a,b}(i-2,j-2)+1 & if\ i,j > 1\ and\ a[i] = b[j-1] and\ a[-1] = b[j] \end{cases} \quad (1)$$

where $1_{(a_i \neq b_i)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise.

Each recursive call matches one of the cases covered by the Damerau–Levenshtein distance:

- $d_{a,b}(i-1,j)+1$ corresponds to a deletion (from a to b)
- $d_{a,b}(i,j-1)+1$ corresponds to a insertion (from a to b)
- $d_{a,b}(i-1,j-1)+1_{(a_i \neq b_i)}$ corresponds to a match or mismatch, depending on whether the respective symbols are the same.
- $d_{a,b}(i-2,j-2)+1$ corresponds to a transposition between two successive symbols.

The Damerau–Levenshtein distance between a and b is then given by the function value for full strings: $d_{a,b}(|a|,|b|)$ where $i = |a|$ denotes the length of string a and $j = |b|$ is the length of b.

Because the Damerau Levenshtein Distance function calculates how many changes are needed to convert *s1* to *s2*, the value of similarity can be defined as Equation 2 [7].

$$Plagiarized\ Value = \left\{1 - \frac{diff}{Max(SS,TS)}\right\} * 100 \quad (2)$$

3. Results and Discussion

The experimental scenario is divided into two. The first is to compare the similarity values obtained by the system when using Damerau Levenshtein and Levenshtein algorithm. Second is testing to find the effect of query expansion on system design.

From the results of the experiments carried out. First, trials on documents that have been deducted from the original content (doc-20% and doc-40%), the percentage of similarity obtained was

ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

low. This is reasonable because the system design created is a word match where the system recognizes the similarity of documents based on the appearance of the same word between 2 documents.

Second, trials on documents that have a different sequence from the original document (doc-SE), the system can recognize the sentences identified by plagiarism even though the location of the sentence is changed randomly. This is indicated by the acquisition of 100% similarity value.

Third, trials on documents which verb structure is replaced to be active from passive or vice versa (doc-AP), the system can also recognize the plagiarism sentences from the two documents that are compared. In this trial, the influence of the stemming process is very important, where the stemming process will get the root words of the entire sentence, so that all kinds of word additions or affix that form active or passive verbs will disappear.

Fourth, trials on documents which words is turned into synonim that have same meaning from the original document (doc-syn). The effect of query expansion greatly affects the value of similarities obtained. Query expansion links vocabulary gaps between documents. The word owned by the test document actually has the same meaning as the original document. The proposed system design in this research can recognize the document even though it is different in content, but it is same in context.

The accuracy of document similarity detection using the Damerau Levenshtein or the Levenshtein algorithm does not give a significant effect. On the fifth trials, on documents with typographic errors from the original document or misspelling of words (doc-typo), the Damerau Levenshtein Distance algorithm gives better similarity results because the transposition operation can handle the exchange of adjacent letter positions. So that the obtained edit distance is more optimal. All experimental results can be seen in Figure 1 and Figure 2, where Figure 1 is a system design without using query expansion, while Figure 2 is a system design by adding query expansion.

Figure 1 Experiments Results Diagram (Without Query Expansion)

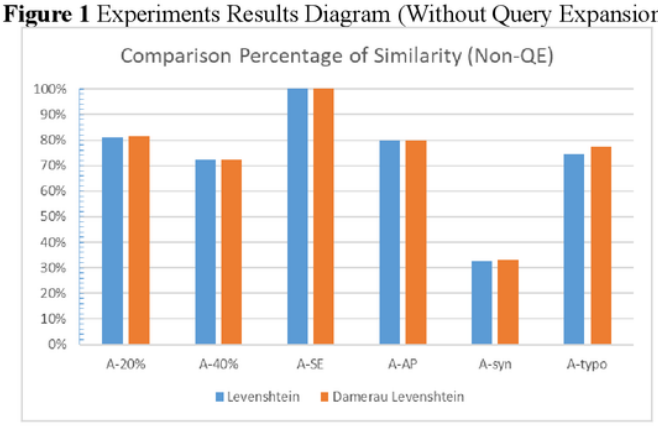
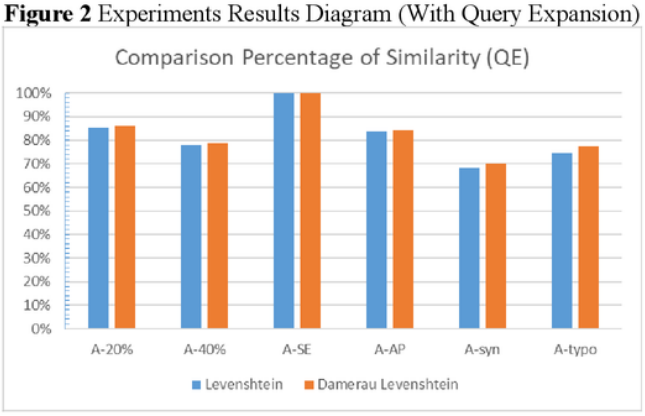


Figure 2 Experiments Results Diagram (With Query Expansion)



ISSN (Cetak) 2527-6042
eISSN (Online) 2527-6050

4. Conclusion

Based on the trial and results analysis, some conclusions can be drawn, including:

1. The application of the Damerau Levenshtein algorithm in detecting text plagiarism is done by using word-for-word matching. Similarity calculation in determining the value of plagiarism by calculating the number of difference characters of the string being compared.
2. The use of stemming produces better accuracy values especially for documents which verb structure is replaced to be active from passive or vice versa.
3. Comparison of the distance between Damerau Levenshtein and Levenshtein algorithm in terms of similarity can be said to be the same as average, but in the case of typography error document testing, the Damerau Levenshtein can recognize plagiarism documents better indicated by a higher similarity value.
4. Query expansion can increase the similarity value by detecting synonyms of words, so the application can detect document plagiarism more quieter well.

It is expected that the further development of this application can recognize the image of documents, so it can detect for entire documents

References

[1] Nurhayati, Busman, "Development of document plagiarism detection software using levensthein distance algorithm on Android smartphone". Cyber and IT Service Management (CITSM), 2017 5th International Conference on, Pp. 1-6.
[2] Christanti, Viny M., and Dali S. Naga. "Fast and Accurate Spelling Correction Using Trie and Damerau-levenshtein Distance Bigram." TELKOMNIKA, Vol. 16, No. 2, Pp. 827-833, 2018.
[3] Pasca, Marius A., and Sandra M. Harabagiu. "High performance question/answering." Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2001.
[4] Sastroasmoro, Sudigdo. "Beberapa Catatan Tentang Plagiarisme." Majalah Kedokteran Indonesia, Vol. 57, No.8, Pp. 239-244, 2007.
[5] Lanin, I. (2009). Kateglo. Retrieved 2015, from https://ivanlanin.wordpress.com/2009/06/11/kateglo/
[6] Frederick J Damerau. A technique for computer detection and correction of spelling errors. Communications of the ACM. Vol. 7, No. 3, Pp. 171-176, 1964.
[7] Su, Zhan, et al. "Plagiarism detection using the Levenshtein distance and Smith-Waterman algorithm." Innovative Computing Information and Control, 2008. ICICIC'08. 3rd International Conference on. IEEE, 2008.

AA11

ORIGINALITY REPORT

10%	12%	5%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to CSU, San Jose State University Student Paper	5%
2	Submitted to University of Newcastle upon Tyne Student Paper	3%
3	en.wikipedia.org Internet Source	2%
4	eprints.umm.ac.id Internet Source	2%

Exclude quotes	Off	Exclude matches	< 2%
Exclude bibliography	Off		